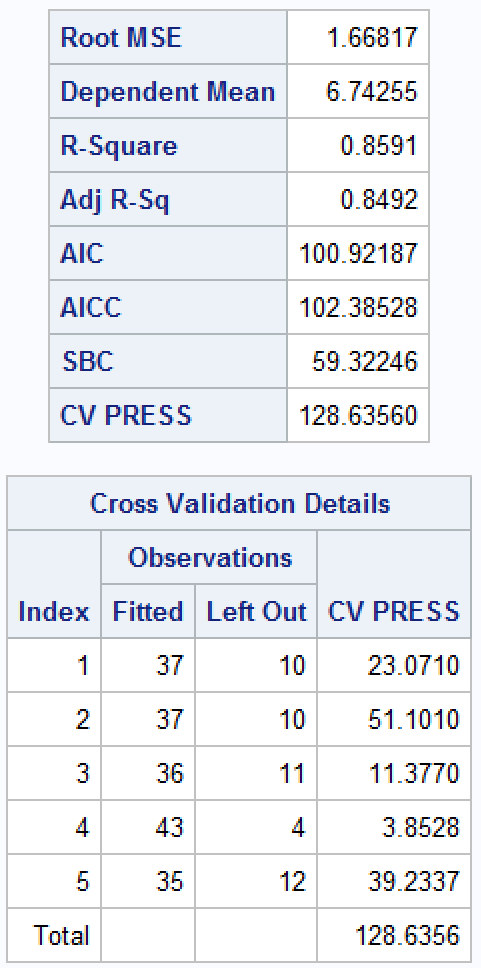
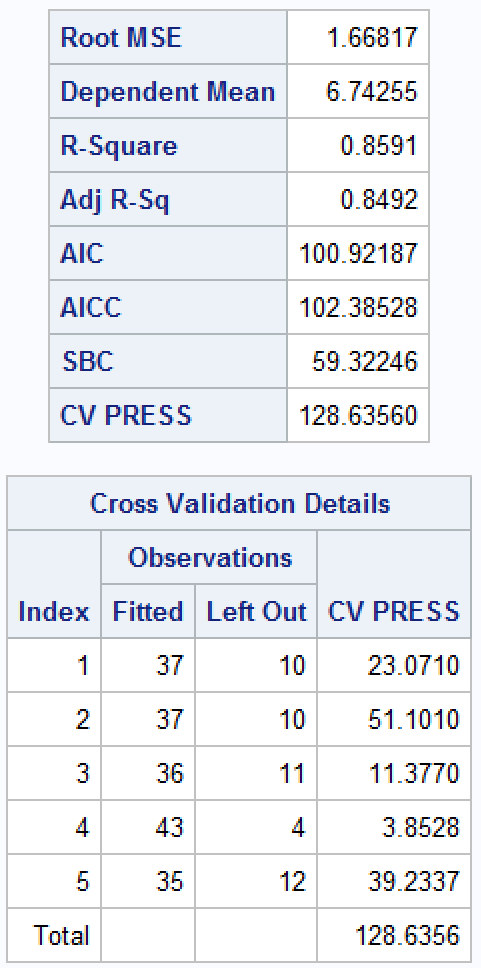
HW Unit 9

Consider data in file hw9CourseData.xlsx. This data from Chicago area zip codes in the 1970s is described below:

* + General Zip Code Features
    - Fire = fires / 1,000 households
    - Theft = thefts / 1,000 population
    - Age = percentage of housing units built prior to 1940
    - Income = median family income
    - Race = percentage minority
    - Zip = zip code
  + Insurance Companies New Policies
    - Vol = number of voluntary policies issued by insurance companies / 100 households
    - Invol = number of involuntary policies issued by insurance companies / 100 households

Run a principal components analysis in SAS on this data with the goal of using the components to understand the effect of the variables on the insurance companies’ voluntary policies. Use the variance/covariance matrix of the variables when calculating eigenvectors and values.

1. Use proc glmselect to regress the voluntary insurance sales on PCs 1 – 5 again. Use a stepwise regression with the select = CV, choose = CV and stop = AIC. Report the SBC and CVPress from the selected model.

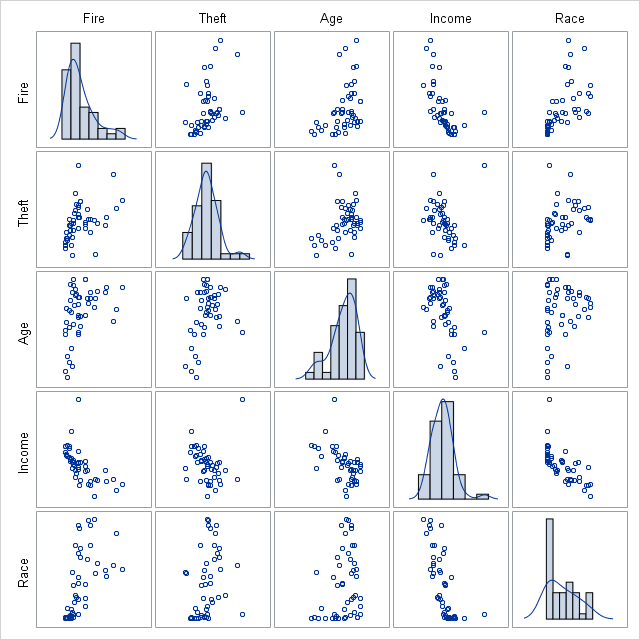
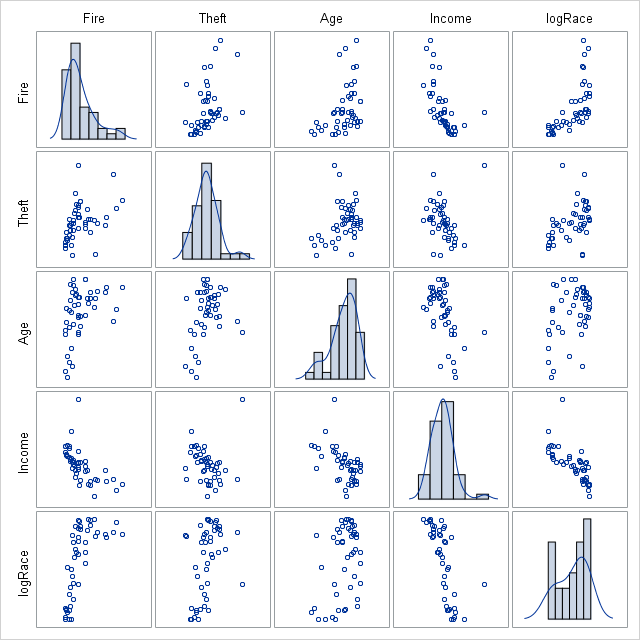


2. Create and display a matrix of scatterplots here for all the variables in the data set with histograms down the diagonal. Transform the Race percent to log(Race). Provide an additional matrix of scatterplots here for all the variables in the data (with log(race) this time) with histograms down the diagonal. Below is the code to do this.

**proc** **sgscatter** data=q;

matrix Fire Theft Age Income Race/diagonal=(histogram kernel);

**run**;



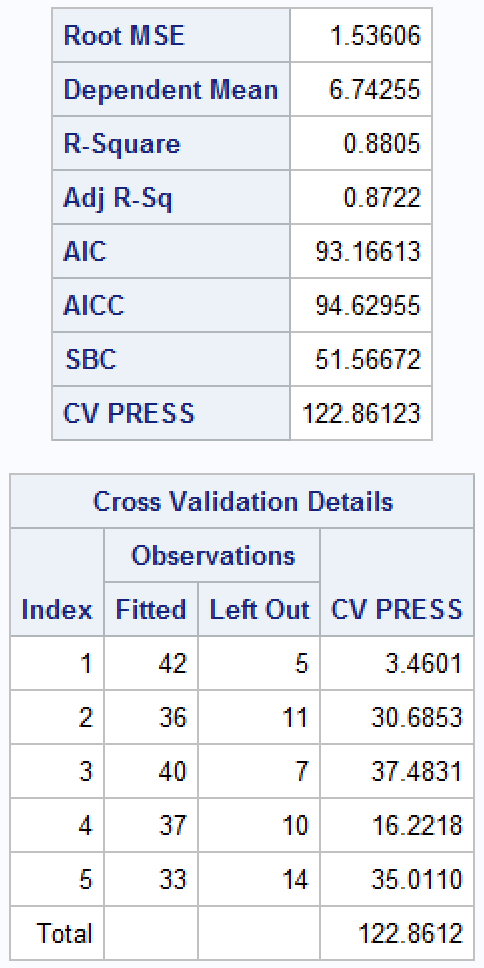
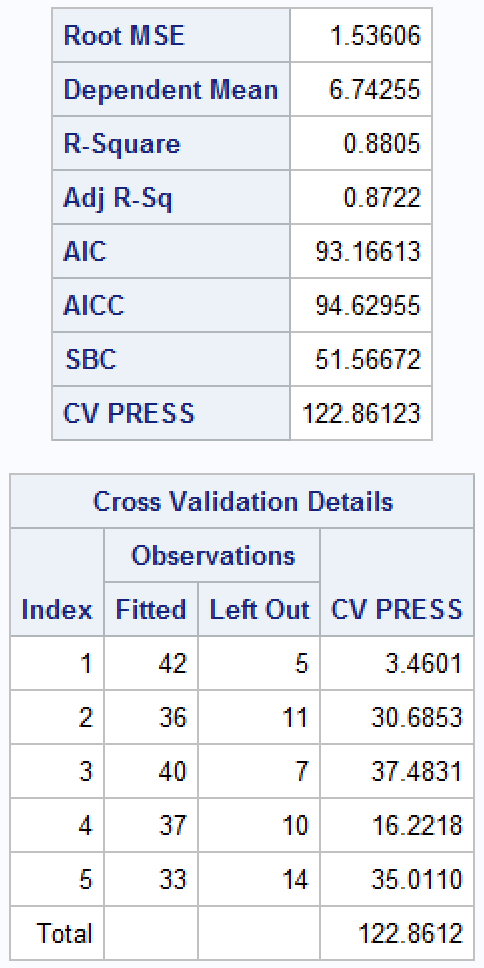
3. Why did we perform the log transform?

A log transformation was performed because the Race variable was right skewed. Normalization is required as an assumption for regression.

4. Did the log transform help?

It appears a log transformation helped with the skew and normalized the distribution.

5. Re-conduct the above model selection procedure in question 1, this time with the log of the race percent. Report the SBC and the CVPRESS.





6. Compare the two models found in question 1 and 5. Which do you prefer and why?

Based on the two model outputs, it appears the model with the log transformation for Race has a better fit. The AIC and SBC is lower in the log transformed model. The CVPRESS is also lower in the second model. We can also see the adjusted R squared is also higher, which supports a greater explanation of the variance.